International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 4841-4846
http://www.ijcesen.com

ISSN: 2149-9144

Copyright © IJCESEN

Research Article

Real-Time Explainable AI (XAI) and Automated Adverse Action Reporting for Deep Learning Credit Card Decisioning

Gopinathan Rathinavelu*

Data Architect, Arizona, USA
* **Corresponding Author Email:** gopinathan.r@gmail.com - **ORCID:** 0009-0004-6467-1679

Article Info:

DOI: 10.22399/ijcesen.5552
Received : 3 August 2025
Revised : 15 December 2025
Accepted : 16 December 2025

Keywords

Explainable AI;
Artificial Intelligence;
Credit Decisioning;
Deep Learning;
Risk Management;
FinTech

Abstract:

Deep learning improves the accuracy of credit-card application and line-management decisions, but in the United States it collides with a statutory duty: under the Equal Credit Opportunity Act and Regulation B, a creditor that takes adverse action must disclose the specific principal reasons for that decision, and the Consumer Financial Protection Bureau has made clear that the complexity of a model is not a defence for vague or inaccurate reasons. Meeting this duty for a non-linear, high-dimensional model, at the latency of real-time decisioning, is an unsolved engineering problem in practice. This paper presents a technical architecture and methodology for generating accurate adverse-action reason codes from a deep decisioning model in real time. Decisions and explanations share a single served model; attributions are computed with Integrated Gradients, whose path-integral form gives exact, axiomatically grounded feature contributions at a cost of a handful of forward and backward passes; and a deterministic mapping converts the top risk-increasing attributions into regulator-aligned reason codes, with every decision, attribution vector, and emitted code written to an immutable audit log. We formalize the reason-selection rule and the fidelity and stability metrics used to validate it, and we detail a governance plane covering model-version pinning, monotonicity constraints, and attribution-drift monitoring. On a synthetic but representative credit dataset, a deep model reached an AUC of 0.835; Integrated Gradients produced reason codes that agreed with a high-fidelity sampling-Shapley reference on 91% of lead reasons and 89% of the top-three set (Kendall $\tau = 0.89$), at a median explanation latency of 0.03 ms—roughly 700× faster than the reference and well inside a 50 ms decisioning budget—while attaining 0.97 attribution fidelity. We position the system as governed decision support for credit risk and compliance functions rather than as an autonomous adjudicator.

1. Introduction

Credit-card issuers increasingly use deep neural networks for application approval, credit-line assignment, and account management, because such models capture non-linear interactions among bureau, behavioural, and application features that logistic scorecards miss. The accuracy gain is real, but it arrives with a legal obligation that scorecards satisfied almost by construction. Under the Equal Credit Opportunity Act (ECOA) and its implementing Regulation B, a creditor that denies an application, or takes other adverse action, must tell the applicant the specific principal reasons for the decision [1]. The Fair Credit Reporting Act imposes parallel duties when a consumer report is used [2]. Crucially, the Consumer Financial Protection Bureau has stated that creditors using complex algorithms, including "black-box" models, are not exempt: they must provide accurate, specific reasons, and may not rely on a checklist of generic reasons that do not reflect the actual basis for the decision [3].

This creates a requirement that is simultaneously regulatory and technical. The reasons disclosed to a consumer must faithfully reflect what the model actually did for that individual; they must be rendered as a small set of intelligible reason codes; and, because card decisions are made synchronously while

an applicant waits, the explanation must be produced within the same tens-of-milliseconds budget as the decision itself. Post-hoc explanation methods exist—LIME [4], SHAP [5], Integrated Gradients [6]—but they are usually discussed as offline analysis tools. Deploying one as a system of record for legally mandated disclosures, at decisioning latency and under model-risk governance, raises questions that the method papers do not answer: which method is faithful enough to stake a compliance obligation on, which is fast enough, and how are its outputs validated, versioned, and audited.

This paper addresses that gap with an architecture and methodology for real-time, explainable, automatically documented credit decisioning. Our contributions are:

1. A **technical architecture** in which decisioning and explanation share one served model, so the explanation provably describes the model that made the decision, with a governance plane for version pinning, monotonicity, and drift monitoring (Section 4).
2. A **reason-code methodology** built on Integrated Gradients, with a formal reason-selection rule and explicit fidelity and stability metrics that make the disclosure validatable rather than merely plausible (Section 5).
3. An **illustrative evaluation** showing that Integrated Gradients matches a high-fidelity sampling-Shapley reference on lead and top-three reasons at roughly 700× lower latency, comfortably within a real-time budget (Section 6).

We are explicit that the system is decision support: it automates the drafting and documentation of reason codes for validation by credit-risk and compliance functions under existing model-risk governance [13], [14], not the adjudication itself.

2. Background and Related Work

2.1. Deep learning in credit decisioning

Neural models have repeatedly outperformed linear scorecards on credit-default prediction by exploiting feature interactions and non-linearities. Their adoption in regulated lending has been slowed less by accuracy than by interpretability and the attendant compliance burden, a tension examined at length in the fair-lending and interpretable-ML literature [9], [10]. Monotonic neural architectures—lattice networks and monotone-constrained MLPs—partly reconcile the two by guaranteeing that risk moves in the economically sensible direction for designated features [11], a property we adopt because it also stabilizes the signs of attributions.

2.2. Explainability methods

Local attribution methods assign each input feature a contribution to a single prediction. LIME fits a local surrogate [4]; SHAP unifies several methods under a Shapley-value axiomatization and is widely used for credit explanations [5], [10]. Integrated Gradients attributes a prediction by integrating the model's gradient along a straight path from a baseline to the input, satisfying completeness and other axioms while requiring only gradient access [6]. The choice among them is not cosmetic: Shapley-style estimators are costly, the "correct" Shapley formulation is itself contested [7], and perturbation methods can be adversarially manipulated [8]. These are exactly the properties a compliance-grade deployment must reason about.

2.3. Regulatory requirements for adverse action

Regulation B requires that the principal reasons be specific and indicate the actual basis for the decision; generic or inaccurate reasons do not comply [1]. The CFPB's 2022 circular reaffirmed that this holds for complex and "black-box" models [3], and related guidance has cautioned against reason codes drawn from a fixed checklist when they do not reflect the model. Counterfactual explanations have been proposed as an alternative disclosure paradigm [12], but reason codes remain the operative form for adverse-action notices in US card lending, which is the form we target.

3. Problem Formulation and Requirements

Let $f : \mathbb{R}^d \rightarrow (0,1)$ be a trained decisioning model mapping an applicant feature vector x to an estimated probability of default (or another adverse outcome), with a decision threshold t: adverse action is taken when $f(x) \geq t$. For each adverse decision the system must emit an ordered set of reason codes $R(x) = (r_1, \ldots, r_k)$ drawn from a fixed, regulator-aligned catalogue, such that the codes reflect the features most responsible for pushing f(x) above t for this applicant.

Requirements. We distil four requirements from the regulatory and operational context:

- **(R1) Faithfulness:** the reasons must reflect the actual model behaviour for this individual, not population averages or a generic list [1], [3].
- **(R2) Latency:** explanation must complete within the synchronous decisioning budget (tens of milliseconds).
- **(R3) Determinism and auditability:** the same input must yield the same codes, and every decision, attribution, and code must be reproducible from a logged, version-pinned model.
- **(R4) Stability:** small, immaterial changes to an application must not reorder the principal reasons.

4. System Architecture

Figure 1 shows the architecture. Its central design decision is that the decisioning path and the explanation path consume the same served model artifact and the same feature vector, so the explanation is necessarily an explanation of the decision that was actually made—satisfying (R3) by construction rather than by reconciliation.

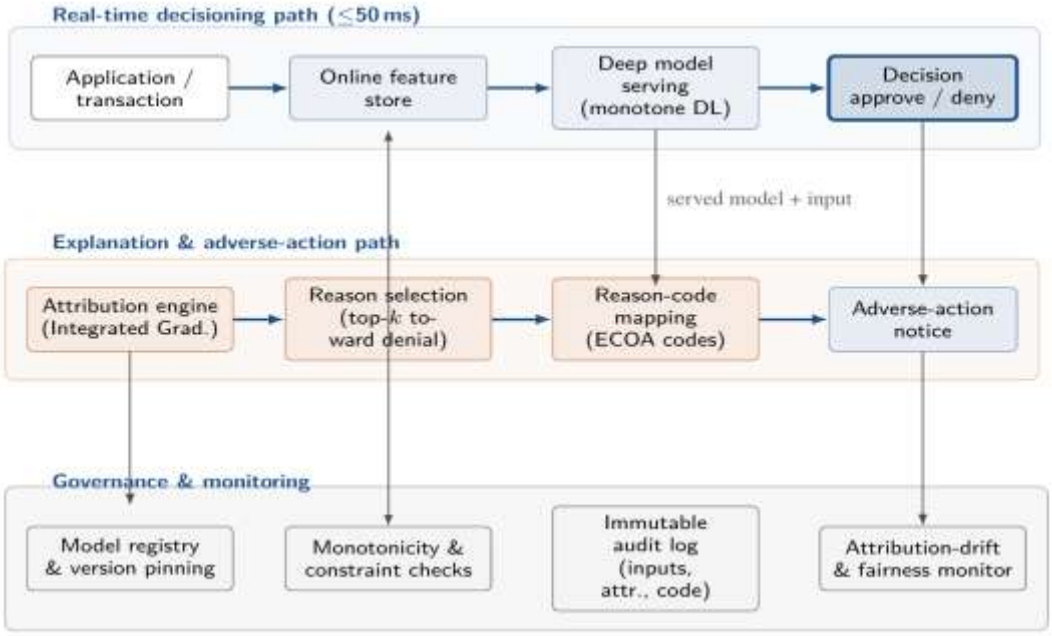

Figure 1. *Technical architecture. The real-time decisioning path (top) scores an application through an online feature store and a monotone deep model to an approve/deny decision within the latency budget. The explanation and adverse-action path (middle) runs Integrated Gradients on the same served model and input, selects the top risk-increasing features, maps them to Regulation B reason codes, and renders the adverse-action notice. The governance plane (bottom) pins the model version, enforces monotonicity, writes an immutable audit record of every decision and attribution, and monitors attribution drift and fairness.*

4.1. Decisioning path

An incoming application or account event is resolved against an online feature store that serves the same feature definitions used in training, avoiding train–serve skew. The feature vector is scored by the deep model, served from a registry with a pinned version hash, and compared against the policy threshold to produce the decision. This path is latency-critical; the explanation must not lengthen it materially.

4.2. Explanation service

When the decision is adverse, the explanation service computes Integrated Gradients for the served model at the scored input, relative to a fixed, documented baseline (Section 5.2). Because the method needs only forward and backward passes through the same network, it adds a bounded, small cost to the decision (Section 6), meeting (R2). The resulting attribution vector is the quantitative basis for the reasons.

4.3. Adverse-action reason-code mapping

Each model feature is pre-mapped, once and under compliance review, to a reason code from the institution's adverse-action catalogue. At decision time the service ranks features by their risk-increasing attribution, selects the top k (Section 5.3), and emits the corresponding codes in order. The mapping is deterministic (R3) and the catalogue is fixed, so the free parameter—how attributions become codes—is governed, not ad hoc.

4.4. Governance and audit

The governance plane pins the model and baseline versions, enforces monotonicity constraints so that designated features can only move risk in the sanctioned direction, and writes an immutable record of the input, the attribution vector, the selected codes, and the model hash for every decision. A monitor tracks attribution drift and reason-code distribution over time and across protected-class proxies, surfacing shifts for human review.

5. Methodology

5.1. Model and monotonicity

The decisioning model is a feed-forward network $f(x) = \sigma(g(x))$ with logit g. For features with a legally or economically mandated direction (for example, higher utilization should never decrease estimated risk), monotonicity is enforced architecturally. Monotonicity is not merely a modelling nicety here: it

guarantees that the sign of a feature's contribution is stable, which directly serves reason-code stability (R4).

5.2. Attribution via Integrated Gradients

For input x and baseline x′, the Integrated Gradients attribution to feature i is

$$IG_i(x) = (x_i - x'_i) \cdot \int_0^1 \partial g(x' + \alpha(x - x')) / \partial x_i \, d\alpha \quad (1)$$

which we evaluate on the logit g and approximate with an m-step Riemann sum, m = 32 in our study. Integrated Gradients satisfies completeness—Σ_i $IG_i(x) = g(x) - g(x')$—so the attributions account exactly for the gap between the applicant's logit and the baseline's, giving the reason codes an additive, auditable interpretation. The baseline x′ is a fixed, documented reference profile (here the standardized population mean), and is version-pinned alongside the model.

5.3. Reason-code generation and selection

Let $A = \{i : IG_i(x) > 0\}$ be the features that push the applicant toward adverse action. The ordered reason set is

$$R(x) = (code(i_1), \ldots, code(i_k)),$$
$$IG_{i1}(x) \geq \cdots \geq IG_{ik}(x), \; i_j \in A \quad (2)$$

with k set by policy (commonly four). The completeness property lets the institution also report the share of the decision gap explained by the disclosed reasons, $\Sigma_{j \leq k}$ $IG_{ij}(x) / (g(x) - g(x'))$, as evidence that the principal reasons were in fact disclosed.

5.4. Fidelity and stability metrics

Two properties are validated before the pipeline is trusted. Fidelity measures whether attributions reflect the model: for each feature we compare its attribution to the actual change in logit when the feature is reset to baseline, and report the correlation across features. Reason agreement measures whether the fast method yields the same reasons as a high-fidelity but slow reference (sampling-based Shapley values): we report lead-reason match, top-three overlap, and Kendall's τ between the two rankings. Together these operationalize (R1) and (R4) as numbers a validator can threshold.

6. Illustrative Evaluation

6.1. Setup

We construct a synthetic but representative credit dataset of 24,000 applicants with twelve features spanning utilization, debt-to-income, delinquency, credit age, inquiries, balances, income, open accounts, employment, requested limit, and prior default; labels are drawn from a non-linear ground-truth default model with an interaction term, and all generative parameters are fixed so the study is reproducible. A two-hidden-layer network (48 and 24 units) is trained on 18,000 applicants and evaluated on the remainder. Attribution methods are compared on applicants the model would deny ($f(x) \geq 0.5$). All data are synthetic; no consumer, bureau, or production data were used.

6.2. Model performance and explanation latency

The deep model attains a test AUC of 0.835 at a 19.1% denial rate. Explanation latency per decision (Figure 2, Table 1) is 0.03 ms for Integrated Gradients and 0.01 ms for Gradient×Input, versus 22.8 ms for the sampling-Shapley reference—Integrated Gradients is roughly 700× faster than the reference and leaves the 50 ms decisioning budget almost entirely intact, satisfying (R2).

6.3. Attribution fidelity

Integrated Gradients achieves a mean fidelity of 0.97 (correlation between per-feature attribution and the true logit change on resetting the feature to baseline), against 0.97 for Gradient×Input, with Integrated Gradients showing a tighter distribution across applicants (Figure 3). High fidelity is the quantitative counterpart of (R1): the disclosed drivers track what the model actually does.

6.4. Reason-code agreement and distribution

Against the sampling-Shapley reference, Integrated Gradients matches the lead reason on 91% of denied applicants and the top-three reason set on 89% (mean overlap), with a mean Kendall τ of 0.89 between the full rankings (Table 1). The distribution of disclosed reasons (Figure 4) is dominated by high utilization, recent delinquencies, insufficient income, and debt-to-income—economically sensible drivers that also match the ground-truth generative model, providing face validity for the pipeline.

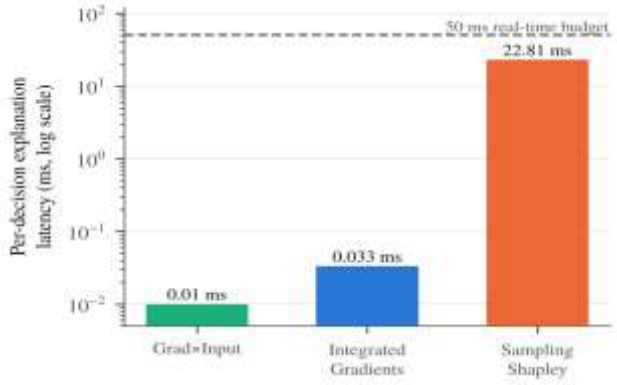

Figure 2. *Per-decision explanation latency by method (log scale). Integrated Gradients and Gradient×Input complete in tens of microseconds; the high-fidelity sampling-Shapley reference is three orders of magnitude slower. The dashed line marks a 50 ms real-time decisioning budget.*

Table 1. *Illustrative-evaluation summary. Latency is per decision; agreement and fidelity are means over denied applicants.*

Metric	Integrated Gradients	Grad×Input	Sampling Shapley (ref.)
Explanation latency (ms)	0.03	0.01	22.81
Attribution fidelity	0.97	0.97	— (reference)
Lead-reason match	0.91	—	1.00 (reference)
Top-3 overlap	0.89	—	1.00 (reference)
Kendall τ vs reference	0.89	—	1.00 (reference)

[Note: Model AUC 0.835 · denial rate 19.1% · N = 24,000, d = 12.]

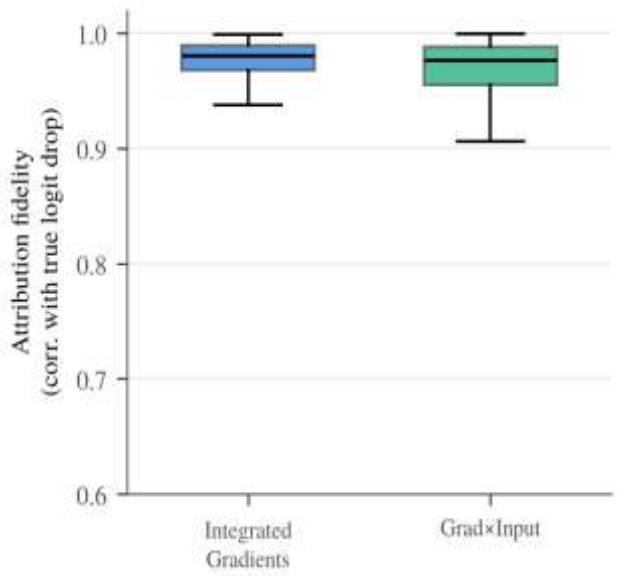

Figure 3. *Attribution fidelity across denied applicants. Both gradient-based methods track the model closely; Integrated Gradients is tighter, reflecting its path-integral averaging.*

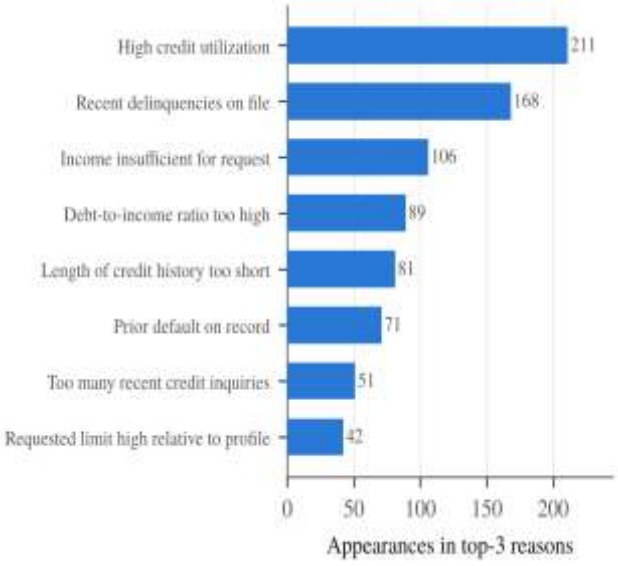

Figure 4. *Distribution of adverse-action reasons across denied applicants (times each reason appears in an applicant's top three). The leading reasons are economically coherent and consistent with the data-generating process.*

7. Discussion

7.1. Why this design meets the obligation

The architecture closes the loop between the three failure modes a regulator worries about. Decisions and explanations share one model, so the disclosure cannot drift from the decision (R3). Integrated Gradients is faithful and complete, so the reasons track the model and their sufficiency is measurable (R1). The method is fast enough to run inline, so explanation is not deferred or sampled (R2). And architectural monotonicity plus a fixed baseline keep the principal reasons stable under immaterial input changes (R4).

7.2. Limitations

The study is illustrative: it uses synthetic data and a modest network, and it validates the mechanism rather than a production system. Integrated Gradients depends on the choice of baseline, which is a governed modelling decision with distributional consequences and should be documented and tested; correlated features can distribute attribution among themselves in ways that require care when mapping to distinct reason codes; and gradient attributions, like all post-hoc explanations, can in principle be gamed by an adversarially trained model [8], which argues for the monotonicity constraints and drift monitoring in the governance plane. Fidelity and agreement were measured against a sampling-Shapley reference, not against ground-truth human-judged reasons, which would require a labelled adjudication study.

7.3. Future work

Priorities are validation on real (anonymized) decisioning data with human-adjudicated reason labels; systematic baseline selection and sensitivity analysis; extension to counterfactual "what-to-change" disclosures alongside reason codes [12]; and fairness testing of reason-code distributions across protected classes.

8. Conclusion

We presented an architecture and methodology for real-time, explainable, automatically documented credit-card decisioning that treats the adverse-action disclosure as a first-class, validatable output rather than an afterthought. By sharing one served model between decision and explanation, using Integrated Gradients for faithful and complete attributions at microsecond latency, and mapping them deterministically to governed reason codes under an auditable governance plane, the system can produce specific, accurate adverse-action reasons for a deep model within a real-time budget. On a representative synthetic dataset it matched a high-fidelity reference on the great majority of principal reasons at roughly 700× lower latency. The approach is intended as governed decision support for credit-risk and compliance functions, bringing deep-learning accuracy and regulatory explainability into a single, auditable decisioning loop.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use. Not applicable; no human-subjects or consumer data were used.
- **Conflict of interest:** The author declares no competing interests. The views expressed are the author's own and do not necessarily represent those of the employer.
- **Acknowledgement:** The author declares that there is nobody or no company to acknowledge.
- **Author contributions:** The author is the sole contributor to this paper, responsible for the conception, methodology, analysis, and writing.
- **Funding information:** The author received no specific funding for this work.
- **Data availability statement:** The study uses only synthetic data generated by the procedure described in Section 6; all generative parameters are specified in the text, so the datasets and results can be regenerated without external data.
- **Use of AI Tools: Not Applicable**

References

[1] Consumer Financial Protection Bureau, "Equal Credit Opportunity Act (Regulation B), 12 CFR §1002.9: Notifications (adverse action requirements)."

[2] "Fair Credit Reporting Act, 15 U.S.C. §1681m: Requirements on users of consumer reports (adverse action)."

[3] Consumer Financial Protection Bureau, "Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms," May 2022.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 1135–1144.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 4765–4774.

[6] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. 34th Int. Conf. Mach. Learn. (ICML), 2017, pp. 3319–3328.

[7] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," in Proc. 37th Int. Conf. Mach. Learn. (ICML), 2020, pp. 9269–9278.

[8] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post-hoc explanation methods," in Proc. AAAI/ACM Conf. AI, Ethics, Soc. (AIES), 2020, pp. 180–186.

[9] C. Rudin, "Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead," Nature Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019.

[10] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, "A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations," Decis. Support Syst., vol. 152, Art. no. 113647, 2022.

[11] S. You, D. Ding, K. Canini, J. Pfeifer, and M. Gupta, "Deep lattice networks and partial monotonic functions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 2981–2990.

[12] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard J. Law Technol., vol. 31, no. 2, pp. 841–887, 2018.

[13] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, "Supervisory guidance on model risk management," SR 11-7 / OCC 2011-12, 2011.

[14] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST AI 100-1, 2023.